INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

Project IST-2001-33562 MoWGLI

Report n. D3.a
Metadata for Mathematical Libraries

Main Author:
George Goguadze

Project Acronym: MoWGLI
Project full title: Mathematics On the Web: Get it by Logic and Interfaces
Proposal/Contract no.: IST-2001-33562 MoWGLI

Contents

1 Introduction

The mathematical knowledge is supposed to have its own "mathematical" structure – the way of organization of knowledge in hierarchies of formal theories. This structure is, however, not always useful for mathematical applications. The applications MoWGLI partners work on are ranging from formal proofs to the publication of informal mathematical texts and eLearning applications. Each of these applications has its own library of mathematical content. In order to take advantage of each other's repositories and knowledge, the common knowledge representation has to be found and the knowledge has to be structured and annotated in a way that it could be used by common tools and services.

This report discusses an important part of the common knowledge representation – the metadata annotations that can serve the tools that will be developed in MoWGLI such as search and retrieval tools, dictionaries, parts of authoring tools, editors.

Choosing RDF as the representation format has an advantage of also employing those tools developed world-wide that are based on Web-standards.

Some of the questions essential for the knowledge representation in MoWGLI and particularly for the representation of metadata are :

- What metadata do MoWGLI applications need in common?
- Can the part of these metadata be used by web-applications, external for MoWGLI?
- Do the libraries of mathematical knowledge possess the fixed structure or can this structure be changed by providing new metadata, oriented to yet another usage of the content?

We shall try answer these question and to estimate the reasonable amount of metadata annotations for MoWGLI and classify them into categories depending on the usage criteria.

2 Definition and Function of Mathematical Metadata

The notion of metadata is well understood – it is "data about data", but we shall fix the concrete definition for the MoWGLI project.

`Metadata is the set of annotations that serve to facilitate the administration of the libraries of mathematical knowledge, the search and retrieval of mathematical knowledge and the reuse of the knowledge by different mathematical applications.`

The metadata can be assigned to individual items as well as to their collections and to the libraries themselves. A certain *inheritance* mechanism for metadata can be defined for the collections of items.

2.1 Structure and Metadata

The knowledge representation in MoWGLI is based on `OMDoc` – the semantic markup language for mathematical documents (see [9]). The knowledge is stored in items that are the atomic pieces of mathematical knowledge with a particular type assigned to them. The typed mathematical items and the semantic connections between them build a content based ontology of mathematical knowledge.

`OMDoc` possesses the mechanism of establishing semantic connections between items of mathematical knowledge. The structure element `theory` of `OMDoc` models the concept of a formal theory. By importing mathematical theories and mapping their semantic constructions one can build the hierarchies of mathematical theories (see [8]). The microstructure of some mathematical items can be complex (proofs, inductive definitions) and affects the macrostructure of the embracing theory.

The metadata can partially describe the structure of formal mathematical ontology. This part of metadata can be generated automatically and serve formal mathematical applications. There are, however, other issues mathematical metadata is concerned with. Namely, to group collections of mathematical items, creating content repositories according to some conceptual (rather than formal) criteria. Examples of such collections are the packages of content for an eLearning application or a library of mathematical publications.

A *collection* is somewhat orthogonal to the library. It can be a subset of the library, but can also unite the items from different libraries according to some conceptual criteria. The items of the same library can form different collections depending on the application using this library. This depends on the way the metadata is assigned to the items of the library. If the library has a static metadata set assigned to its elements then the different applications can only generate different views of the same content collections. But as soon as new metadata is assigned to the elements of the library, the new collections can be created.

2.2 Purpose of Metadata

So, what are the purposes of metadata annotations for mathematical content?

Metadata is assigning some properties to the items and the way of the organization of these items depending on the goals of the applications using the mathematical knowledge. This means, it does not only reflect the mathematical structure of the content, but annotates them with additional information, needed in order to facilitate the management of the content. The pure mathematical structure of the connections between items expressed by the constructions of `OMDoc` is in some cases too informative for the purposes of many applications.

Facilities enjoyable by all applications are, for example, semantic search and retrieval of mathematical knowledge.

Searching for a theorem justifying a definition, or the retrieval of the theorem needed for a proof of another theorem can be useful for formal application such as COQ[1] as well as for a tutoring system. Possessing reusability of the content is an enormous facility for a user-adaptive learning environment such as ActiveMath[2].

Basic administration of the content, lifecycle of the documents and other technical metadata as well as the information on the copyright of the content are important for all applications.

3 Collecting Experience from Other Sources

There exists several well-established Web standards for metadata. These standards differ in their purposes and user categories.

We discuss the functionalities of these standards and employ the metadata appropriate for our purpose. Apart from this, we explore several existing projects dealing with mathematical knowledge representation, their usage of standards and own refinements of these standards.

3.1 Metadata Standards

The most established Web standard for basic administrative metadata is Dublin Core Metadata Element Set ([5]). It defines the general metadata most of the Web resources can be annotated. These are `Title`, `Creator`, `Subject`, `Description`, `Publisher`, `Contributor`, `Date`, `Type`, `Format`, `Identifier`, `Source`, `Language`, `Relation`, `Coverage`, and `Rights`.

For some of these elements Dublin Core introduces further refinements using so called qualifiers. For example an element Relation can have a type 'Requires' or 'Is Required By', 'References' or 'Is Referenced By' etc.

Apart from qualifiers Dublin Core Metadata Initiative defines the recommended vocabularies to be used for the values of some of the metadata elements. For example it suggests the values for the `Type` element that describes the type of the digital resource. Among the values from the vocabulary are, for instance 'text', 'dataset', 'event', 'image', etc.

Dublin Core provides guidelines for expressing its elements in HTML, XML and RDF. Several other standards use Dublin Core as basics and introduce their refinements and extensions.

[1] `http://coq.inria.fr/`
[2] `http://www.activemath.org`

IEEE standard for Learning Object Metadata (LOM) [4] is supposed to annotate mainly so-called learning objects that are the information entities used in a learning application, but it defines a lot of detailed technical metadata useful not only for learning applications. For example, some additional attention is paid to the lifecycle, versioning and meta-metadata. LOM introduces a number of other refinements of Dublin Core, for instance the "Classification" category and some refinements for copyrights. Naturally, it also has a big "Educational" category.

There exist several variants of IEEE LOM recommendations. Among others, IMS [3] and SCORM [4] are worth to mention.

3.2 Related Projects

There have been several projects dealing with mathematical knowledge representation, that have gained a certain experience in using Dublin Core and LOM and making their own extensions.

One of such projects is a TRIAL SOLUTION[5] - "Tools for Reusable, Integrated, Adaptable Learning - Systems/standards for Open Learning Using Tested, Interoperable Objects and Networking". It is a project funded by the EU as part of its Information Society Technologies Program (IST), which is a major theme of research and technological development within the EU's Fifth RTD Framework Program.

TRIAL-SOLUTION developed their own metadata, using the elements of DC and IMS as a kernel and defining their own extension.

The knowledge in TRIAL SOLUTION is decomposed in so-called units. The additional elements specific to TRIAL SOLUTION hold the information required to identify single decomposition units to trace the sources of these units, to give details on relations between them and to describe their content using controlled vocabularies.

The keywords in TRIAL SOLUTION have well organized hierarchical structure and form so-called thesauri that facilitates the search and retrieval of the content.

Another interesting project dealing with administration of the libraries of mathematical publications is EULER[6].

EULER is a European based world class real virtual library for mathematics with up-to-date technological solutions, a sound sustainable business model, well accepted by users.

EULER project uses DC elements and qualifiers (refinements), and refines the vocabularies of DC. Among others, they use controlled keyword vocabularies of several Mathematical Classification Systems, refine the DC types of an electronic resource and use several schemes for identifiers. Among the schemes, used in this project are Library of Congress Subject Headings (LCSH), Mathematical Subject Classification (MSC), Deway Decimal Classification (DDC), and Computing Classification System (CCS).

ActiveMath Learning Environment for Mathematics ([11]) uses Dublin Core as general metadata and make several refinements of it.

The relations between items of mathematical knowledge are obtaining more types of mathematical as well as of a pedagogical nature. ActiveMath uses some educational metadata of LOM such as `difficulty` of a resource, the `learning_context` etc., but also defines its own educational metadata extensions such as an `abstractness` of the resource, `competence_level` used to differentiate between the pedagogical that exercise aim at, etc. (see [7])

Another project dealing with libraries of mathematical knowledge is HELM – Hypertextual Electronic Library of Mathematics ([1]). This project is meant to integrate the current tools for the automation of formal reasoning and the mechanization of mathematics (proof assistants and logical frameworks) with the most recent technologies for the development of web applications and electronic publishing.

Currently it uses Dublin Core metadata and makes its own refinements in order to enable semantic queries on the formal content of its library.

[3] `http://www.imsglobal.org/metadata/`
[4] `http://www.adlnet.org/`
[5] http://www.trial-solution.de/
[6] `http://www.emis.de/projects/EULER/`

4 Metadata Representation

Along with the development of the Web languages the tries to encode metadata information in this languages were made. Firstly, the additional annotations of HTML documents were introduced. One of them was of Dublin Core Metadata Initiative.

As the next step, XML was tried out. XML DTD appeared to be too well-founded and not flexible enough for the representation of metadata. XML schema provided more flexibility, but the need of some more scalable and less restrictive representation arose.

As the solution the Resource Description Framework (RDF) was suggested by the W3C community. [7]

4.1 Why RDF

The Resource Description Framework (RDF) is a framework for describing and interchanging metadata. It is a representation format itself and a grammar, allowing to define the languages for more precise description of specific resources (see [12]).

For example, we can refine the notion of the the main Class of RDF `Resource`, by defining the subclass of it called `MathResource`. We can also define `MathItem` and `MathCollection` to be the further subclasses of `MathResource`. The metadata elements we define are the properties of the `MathResource` class and its subclasses.

RDF provides us with the possibility to define languages for semantic annotation of resources.

The RDF approach is similar to Object-Oriented programming. It defines `Classes` of objects and their `Properties`. However, as opposed to Object-Oriented programming the properties are defined by specifying the classes of their domains and ranges rather then classes are described by their properties. Having some specific needs for annotation, one can refine the basic RDF schema and define new Classes and Properties.

The RDF document consists of so-called `Statements` that are triples (Resource, Property, Value).

The RDF annotations are extensible – one can always add more statements. The RDF schemas are easy to refine by just defining new subproperties of existing properties. Annotations are modularized via namespaces. RDF/XML syntax is very general - not as restrictive as DTD. RDF annotations are easy to interchange via their XML representations.

Finally, since RDF statements are just triples they are easy to handle and look things up.

4.2 Structure of Metadata in MoWGLI

According to the functionalities that metadata provides, we classify it in three main categories: **Administrative**, **Mathematical** and **Application-Dependent**.

The first two categories are common for all MoWGLI applications. The **Administrative** metadata can be, in principle usable even for external Web-applications. The **Application-Dependent** metadata is oriented to particular applications of the MoWGLI partners.

First category serves the administration of the resources, versioning, copyrights, and other technical characteristics of the resources. This category of metadata is not particularly specific to mathematical resources.

We divide it conceptually in subcategories **General**, **Lifecycle**, **Technical**, and **Rights**.

General metadata contains basic Dublin Core elements such as `Title`, `Creator`, `Contributor`, `Language` etc. Taking in consideration the best practice recommendations of standards such as Dublin Core and LOM and analyzing the practice of several projects, considered above we select the refinements of the Dublin Core elements used and the vocabularies suitable for the needs of MoWGLI applications.

Second category consists of mathematical metadata such as semantic relations between math items and the keyword annotations using different math classification systems. For mathematical relations, defined in the subcategory **Relation** we refine the Dublin Core `Relation` element by

[7] `http://www.w3.org/RDF`

introducing different kinds of mathematical relations as it is done, for instance, in LOM. We also introduce the `relDirection` property with the help of which we can construct the inverse relation of every kind. In this way, by providing only one relation, one can generate the inverse relation of the same kind pointing to the current item from the item it is related to.

The subcategory **Classification** serves the purpose to annotate the mathematical knowledge with keywords using vocabularies of well-established Mathematical Classification Systems. We follow the practice of the EULER project and select the following systems: Library of Congress Subject Headings LCSH), Mathematical Subject Classification (MSC), Deway Decimal Classification (DDC), and Computing Classification System (CCS).

Third category is due to the specific needs of MoWGLI applications. The metadata defined there is not supposed to be useful for all MoWGLI applications and therefore marked as application-dependent. For instance, educational and formal mathematical applications need more kinds of relations. Educational applications define additional pedagogical metadata, publishing applications need to define more types of mathematical resources such as proceedings, review or thesis that would not be of use for all MoWGLI partners.

The precise definition of the metadata element set and markup for metadata model is given in the next report.

References

[1] Asperti, A., Padovani, L., Sacerdoti Coen, C., Schena, I., "HELM and the semantic Math-Web", Proceedings of the 14th International Conference on Theorem Proving in Higher Order Logics (TPHOLS 2001), 3-6 September 2001, Edinburgh, Scotland.

[2] Berners-Lee, T., "Universal Resource Identifiers in WWW", RFC 1630, CERN, June 1994.

[3] Bray, T., "What is RDF", O'REILY xml.com, January, 2001. http://www.xml.com/pub/a/2001/01/24/rdf.html

[4] Draft Standard for Learning Object Metadata, IEEE P1484.12.2/D1, 2002-09-13, IEEE Learning Technology Standards Committee http://ltsc.ieee.org/wg12/

[5] Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003-02-04 http://www.dublincore.org/documents/dces/

[6] Duval, E., Hodgins, W., Sutton, S., Weibel, S.L., "Metadata Principles and Practicalities", D-Lib Magazine, April 2002, Volume 8, Number 4, ISSN 1082-9873

[7] Goguadze, G., "Knowledge Representation in ActiveMath", SEKI-Report SR-02-02, FR Informatik, Universität des Saarlandes, 2002

[8] Hutter, D., "Reasoning about theories", Deutsches Forschungszentrum für Künstliche Intellligenz (DFKI), 1999

[9] Kohlhase, M., "OMDoc: Towards an OpenMath Representation of Mathematical Documents", Seki Report, FR Informatik, Universität des Saarlandes, 2000.

[10] Kohlhase, M., Franke, A., "MBase: Representing Knowledge and Context for the Integration of Mathematical Software Systems", Journal of Symbolic Computation 23:4 (2001), pp. 365 - 402.

[11] Melis, E., Büdenbender, J., Andres, E., Frischauf, A., Goguadze, G., Libbrecht, P., Pollet, M. and Ullrich, C., ‘‘ActiveMath: A Generic and Adaptive Web-Based Learning Environment’’, Artificial Intelligence and Education, Volume 12, Number 4, 2001.

[12] Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999.
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/